View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 1997 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

December 1997

The Quality of Data and its Effect on Information Usage

Janet Aisbett
The University of Newcastle

Greg Gibbon
The University of Newcastle

Felicity Lear
University of Tasmania

Follow this and additional works at: http://aisel.aisnet.org/pacis1997

Recommended Citation
Aisbett, Janet; Gibbon, Greg; and Lear, Felicity, "The Quality of Data and its Effect on Information Usage" (1997). *PACIS 1997 Proceedings*. 66.
http://aisel.aisnet.org/pacis1997/66

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 1997 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

The Quality of Data and its Effect on Information Usage

*Janet Aisbett, Greg Gibbon and Felicity Lear**
The University of Newcastle and
** University of Tasmania*

mgjea@cc.newcastle.edu.au
mgggg@cc.newcastle.edu.au
Felicity.Lear@infsys.utas.edu.au

Executive summary
The purpose of this paper is to quantify some effects of data quality on the way information can be used.

A restaurant chef does the morning's marketing by selecting produce on the basis of visual presentation, price, and past performance of the supplier. At the same time, his restaurant has posted a range of menus in which both meals and ambience are lavishly described, along with price in small font. But information is not taken at face value by either the chef or the potential diners. Rather, the acceptance of each item is based partly on the decision maker's knowledge of the source and age of the item, and partly on how credible the information seems, that is, how well it accords with a prior world-view. This is true regardless of whether the chef and the potential diners are acquiring the information from the Web or from pedestrian search.

While decision makers in general are sceptical, organisational decision makers accessing electronic information under internal organisational control are not. Instead, they assume a high level of data quality. This assumption cannot be transferred to information drawn from diverse Internet sources. Ideally, therefore, such data providers should attach basic quality labels to published data. Already, multimedia datatypes such as images or spatial data are routinely labelled with source, timestamp and/or precision. However, in most data domains, there is as yet no agreed way to present and estimate data quality. And even if there were quality labelling standards and routine use of labels, data users would still need to interpret labels in the context of their particular task. Moreover, many quality parameters are user-dependent, rather than intrinsic to data.

This paper briefly reviews suggestions for describing the quality of data content. It looks at schemes for dealing with uncertain information including labelling. It then shows quantitatively how data content quality affects two fundamental uses of information, namely, decision making and learning about a subject. The decision making task is to select the "best" choice from a set of alternatives. If data quality is taken into account, the choice may, not surprisingly, be quite different to the choice made by an unsceptical decision maker. When the value of information with respect to a set of interests is defined, its degradation by poor quality can be quantified and so the impact on potential learning can be investigated.

1. The Problem

Organisations which make effective use of the rich information sources on the Internet have a potential competitive advantage. However, information drawn from the unmanaged Web environment is often absorbed by and distributed within organisations without sufficient regard for data quality (Segev 1996). Thus, good decision making may actually be hindered by external electronic access, especially given managers and professionals tend to accept data without questioning its quality (Redman 1995).

Content quality of data in general refers to the degree to which the data represent the "true" state of the world of interest to the recipient of the data. Perfect information provides everything that the recipient wants to know, in a form they can use efficiently. Information that is imperfect in content either leaves uncertainty as to the true state, or worse, provides an erroneous representation of that state. Imperfect content is introduced into information by mechanisms including:

- opinion which may not correctly reflect the state of the world or accord with the opinion of the recipient, eg. a restauranteur's description of a menu;
- data which is a partial representation of the real-world state but may be accepted as complete, eg. extrapolated terrain data leading to erroneous or imprecise elevation information;
- predictions of future states, which necessarily produce uncertain data but may be accepted as fact, eg. forecast of economic indicators;
- incorrect information propagated through ignorance or, occasionally, deliberate misinformation, eg. to boost sales;
- change in the state of the world causing data to no longer represent the current state, eg. a superseded price list.

The effect of these mechanisms is difficult to combat even in corporate data; take for example the vitae of staff members collected for staff development purposes. The situation is far worse when information outside the control of an organisation is used by its decision makers. Decision makers need to be aware of the quality of data in order to properly adjust for uncertainty in content.

2. Alternative approaches to recording uncertainty

There are various strategies for dealing with uncertainty as to the correctness of data. Three common ones, which may be used alone or in concert, are to: annotate data about which there is uncertainty; carry different versions of data, usually with a preferred or master version; label data with quality evaluations.

Annotation in the traditional paper world is generally performed through footnotes drawing attention to alternative versions of "truth". This method is also used in many corporate databases where a text comment field flags alternative possibilities for a datum, or lists other caveats on taking the datum at face value. The limitation of this technique is the need for an experienced user to reliably interpret the comment field. Its advantage is that it allows uncertainty to be represented within a conventional database structure.

Multiple versions of data in Truth Maintenance Systems are maintained so that each is internally logically consistent (de Kleer 1986). However, versions are mutually inconsistent. The master copy is that which is judged to be the best representation of the real world state. New information concerning existing data, or concerning a change in the real world state, may lead to another version being accepted as "best". Truth Maintenance Systems treat data non-uniformly, in that incorrect data are accepted unquestioningly as long as they do not conflict with existing data.

Multiple versions of data are also maintained in systems dealing with engineering or architectural data, with temporal data or with multiple levels of security, and so on. In multilevel secure databases, versions accord with security classification levels, and the master version which best represents the real world is generally that with the highest classification level. In any database supporting multiple versions, the appropriate granularity of versioning is a significant design question: should it be at the attribute level, at the object level, at the table or class level, or at the full database level?

Labelling of data is common with multimedia datatypes such as images. Timestamping and source details are routine in image "headers" or labels, eg. (Flickner et al 1995). Precision labels are routine with geographic data. However, reaching agreed quality metadata for images or for geographical information is a long process (Federal Information Processing Standard 173 1992). Labelling of standard alphanumeric data is less common, in part because of apprehension about implications for storage space and retrieval complexity. Standardisation efforts for such data are directed at developing common terminology for information interchange (eg. ANSI X3L8 Data Representation) rather than at quality metadata.

3. Describing data quality

While data quality metadata definitions remain unstandardised and unagreed (Dvir and Evans 1996), the parameters most commonly used in the literature are:

- accuracy
- reliability
- timeliness
- relevance, and
- completeness.

(Wang and Firth, quoted in Wand and Wang (1994)). There is no consensus as to the definition of these terms. However, the majority of workers agree that *accuracy* refers to correctness of data against the real world they represent. *Reliability* is associated with correctness of the data over time. *Timeliness* (and the closely related concept of *currency*) is the availability of the information "in time" to be of use. *Relevance* is its pertinence to a task. *Completeness* is its ability to represent all that is required of the real world.

Wang, Strong and Guarascio (1994) investigated what data consumers understood by data quality. They clustered the survey results to form four groups, corresponding to qualities intrinsic to the data, dependent on the context, referring to data representation, and referring to accessibility and access security. The first two, content-related, groupings are described as covering the following qualities:

- *intrinsic qualities*: accuracy, objectivity, believability, and reputation
- *contextual qualities:* value-added, relevancy, timeliness, completeness, and appropriate amount.

On the other hand, Wand and Wang (1994) determine what they call an *internal view*, related to design and operation, and an *external view*, related to use and value of information. Qualities were assigned to each as follows:

- *internal view*: accuracy, reliability, timeliness, completeness, currency, consistency, precision
- *external view:* timeliness, relevance, content, importance, sufficiency, usability, usefulness, clarity, conciseness, freedom from bias, informativeness, level of detail, quantitativeness, scope, interpretability, understandability, flexibility, format, efficiency.

These disparate ways of looking at quality reflect the complexity associated with almost all concepts to do with information. There are many perspectives on quality presented in the literature.

4. Quality parameters used by military intelligence analysts

We will restrict attention to quality labelling derived from military intelligence doctrine, since this has a firm historical basis. See for example, the Manual of Land Warfare (1979). It is usual in this context to attribute the concept of *reliability* to a source of information rather than to an information item itself. Reliability connotes accuracy over many releases of information. Thus, information from a retailer's home page may not be considered reliable, whereas a government statistical agency would have a high level of reliability. (This notion may have been what respondents interpreted "reputation" to mean in the survey conducted by Wang et al, 1994).

Reliability of a source is a key quality indicator: if a source is unreliable, the true datum value for information released from that source may lie anywhere in the allowed domain. Military analysts assign discrete reliability codes to data as in Table 1(a).

(a) Reliability of Source	*(b) Accuracy of Information*	*(c) Credibility of Information*
A - completely reliable	**1** - confirmed by other sources	**1** - confirmed
B - usually reliable	**2** - probably true	**2** - most likely
C - fairly reliable	**3** - possibly true	**3** - likely
D - not usually reliable	**4** - doubtful	**4** - probable
E - unreliable	**5** - improbable	**5** - possible
F - reliability cannot be judged	**6** - truth cannot be judged	**6** - unlikely
		7 - refuted

Table 1: Typical military discretisation of source reliability and information accuracy and credibility

If there is opportunity to later assess correctness of data, reliability of a source can be determined statistically. This is done, for example, by Meteorological Bureaux when analysing submitted reports from Weather Recorders.

Military analysts attribute *credibility* to a data item in terms of the believability of its content. The analysis of survey data by Wang, Strong and Guarascio found credibility to be an intrinsic property. However, any estimate of the credibility of an item of information is based on assessment of its coherence with existing model(s) of the world, which are generally subjective. In assessing credibility, military analysts assign a degree of plausibility to a datum, which has to be reasonably high to justify further processing. Estimating credibility in practice requires caution, else important but unexpected information may be dismissed (Manual of Land Warfare 1979). In order to reduce information overload, analysts traditionally assess whether intelligence information is credible or otherwise, and "throw away" the latter. However, rather than this all-or-nothing approach, in computer-supported environments numerical rankings have been suggested, as in Table 1(c).

The *accuracy* of data is the degree to which they accord with the true state of the world that they represent, or what is accepted to be the true state (ISO, ISO/TC 211 1996). *Precision* is variously seen as a related concept or as a subconcept. Accuracy and precision estimates can be assigned by a data source according to the accepted views of the world at that time, but these may not always be the views accepted by the recipient. Accuracy may be evaluated after the event, by comparison with other information. Table 1(b) gives labels used in military assessments of accuracy.

Military information is also *timestamped*, and often has a expiration data to indicate currency.

Two further parameters describing information used in military intelligence will not be used later in this paper but are nevertheless worth noting because they will have increasing importance in Web based information systems. *Confidentiality* indicates the degree to which data needs to protected against unauthorised release. Protecting knowledge of data values, or even knowledge of data existence, may lead to disinformation or to data-hiding behind sanitised versions of data. Confidentiality levels are generally assigned by the originator of data, with protection of data a condition of their (limited) release. The second parameter, *urgency*, is associated with incoming data and has long been associated with information incoming to any organisation. The greater volume of dynamic data brought by the Web raises the importance of priority labelling of information to all organisations.

5. Selecting between alternatives

This section looks at the effect of data content quality on the standard decision task of choosing between a set of alternatives on the basis of preferences for the outcomes. This is the task faced by a restaurant chef who conducts the morning's marketing by selecting a side of beef, say, on the basis of visual presentation, price, and past performance of the supplier. It is also the task faced by browsers deciding which restaurant in a locality to frequent on the basis of a range of menus posted by the restaurants, in which meals and ambience are lavishly described.

Data gathering in either case may take place from the Web, or from traditional pedestrian means. Data is collected against a set of requirements of the decision maker, eg. 'corn-fed beef', 'no more than $3 per kg', 'medium size', or 'Chinese food', 'no MSG', 'less than 5000 calories'. The information so gained will be used to select which carcass from which supplier best meets the restauranteur's requirements, or which restaurant best meets the potential diner's requirements.

A standard decision theory treatment of the decision process is to represent the problem as the choice of the alternative *i* which maximises the utility $U_A(i)$ defined as:

$$U_A(i) = \sum_a u_a(a). \qquad (1)$$

Here, $u_a(a)$ is the utility gained from the outcome *a* attached to satisfaction of the requirement *a* given the choice *i*. The summand is over the set *A* of requirements. See for example Bell (1993).

In the decision situation posed in this paper, the predicated outcome in terms of level of satisfaction of the requirement *a* is estimated using the datum $\underline{x}$ provided by the source about the requirement. Let $a(\underline{x})$ denote the outcome associated with requirement *a* that is determined by the datum $\underline{x}$.

Because low quality information cannot be taken at face value (whether by the chef or the potential diners), the outcome has a probabilistic rather than a deterministic relationship with $\underline{x}$. The nature of this relationship will depend on the source (site) *i*. Let $p_i(\underline{y}|\underline{x})$ be the conditional probability that a real world situation described by $\underline{y}$ is being presented by source *i* as $\underline{x}$.. The decision is then to choose *i* to maximise the expected conditional utility

$$U_A(i, \underline{x}) = \sum_\alpha \sum u_\alpha(a(\underline{y})) \, p_i(\underline{y}|\underline{x}), \qquad (2)$$

where the inner summand is over all valid real world states $\underline{y}$.

6. The effects of content quality on selections between alternatives

This section shows how the data quality parameters considered in section 4 affect decision making. Throughout, $a(\underline{x})$ denotes the outcome associated with *a* determined by the datum $\underline{x}$ if it could be taken at face value. For clarity, let $\underline{x}^d$ denote a value presented by the source as $\underline{x}$ about which there is uncertainty. Any datum $\underline{x}$ can be presented by any information source (site). The datum describes a state of the world which would result from choosing the site which presents it, but this state may not be the true state. Unless otherwise indicated, site *i* is being considered.

First, consider reliability. To each reliability label **A - E** associated with a source *i* in Table 1(a), assign a reliability level $r(s)$ between zero (unreliable) and 1 (completely reliable) according to the estimated probability that information provided by the source represents the true state of the world. If the reliability of this source cannot be judged, then set $r(s)$ to be the mean of the reliability levels over all sources.

Given a datum $\underline{x}^d$ received from source *i* with reliability level $r(s)$, there is probability $r(s)$ of the real world state having correct representation $\underline{x}$, and probability $(1 - r(s))p'(\underline{y})$ of it having correct representation at an arbitrary position $\underline{y}$ in the domain from which the value $\underline{x}$ is drawn. Here, p' is the *a priori* distribution of occurrences of values in that domain. Thus the expected utility for the decision maker derived from the alternative presented by source *i* is

$$\begin{aligned} U_A(i, \underline{x}) &= r(s)\, U_A(i) + (1\text{-}r(s)) \sum_y U_A(i, \underline{y}))p'(\underline{y}) \\ &= r(s) \sum_\alpha u_\alpha(a(\underline{x})) + (1\text{-}r(s)) \sum_\alpha \sum_y u_\alpha(a(\underline{y})) \, p'(\underline{y}). \end{aligned} \qquad (3)$$

If, in the case of certain information, site *i* which published the information $\underline{x}$ was the source with greatest utility, then $\sum_\alpha u_\alpha(a(\underline{x})) > \sum_\alpha u_\alpha(a(\underline{y}))$ for each *y*. Hence the utility provided by this source must be reduced when unreliability is taken into account. The extent of this reduction depends on the utility provided by alternative states of the world; in general, it would be significant.

Credibility is attached to a data item, rather than to the source. So suppose $c(\underline{x}^d)$ is a value between zero and one determined by the credibility assessment of $\underline{x}^d$ in Table 1(c).If the reliability of a source is one, the datum must be assumed correct and formula *(2)* used. If not, the chance that the datum value is $\underline{x}$ equals the product of its credibility and the source reliability; and otherwise it can be distributed anywhere in the valid domain. That is,

$$\begin{aligned} U_A(i, \underline{x}) &= r(s)\, c(\underline{x}^\delta)\; U_A(i) + (1\text{-}r(s)\, c(\underline{x}^\delta)\; \sum_y U_A(i, \underline{y})p'(\underline{y}) \\ &= r(s)\, c(\underline{x}^\delta) \sum_\alpha u_\alpha(a(\underline{x})) + (1\text{-}r(s)\, c(\underline{x}^\delta)) \sum_\alpha \sum_y u_\alpha(a(\underline{y})) \, p'(\underline{y}). \end{aligned} \qquad (4)$$

Again, lack of credibility can affect a decision made on the basis of maximising utility. However, imperfect credibility may not prevent a source from being selected.

How data accuracy is reported depends on the source and, possibly, the actual information item. It may be reported as in Table 1(b), in which case its effect is analogous to that of the previous qualities. However, accuracy may also be reported as a precision, or as a probability distribution p'' of data values. More commonly it is reported via a datum $\underline{x}$ with bounds $\pm e$ on each component of the

vector value, with an implied normal distribution with mean $\underline{x}$ and standard deviation $\underline{e}$. The expected utility gained from choosing the alternative corresponding to the source *i* is therefore

$$U_A(i, \underline{x}^\delta) = \Sigma^{x+e}_{x-e} U_A(i, y)p''(y) = \Sigma_a \Sigma^{x+e}_{x-e} u(a(y))\, p''(y) \qquad (5)$$

If p'' is not provided with the data and is not implied to be normal, it would be estimated by the *a priori* probability p' integrated between the bounds $x \pm e$.

Timeliness is relative to the rate of change of the data. For example, the distribution of turnover of carcasses may be in hours or days (and be limited by the shelf life) whereas a restaurant menu may be static for weeks, or months. The effect of time on the quality of data is complex. For example, a butcher's home page will lose currency most often because it refers to produce no longer available; however, it may also be out-of-date because it refers to freshness of produce still stocked for which there has been gradual deterioration in the quality of the stock. In general, the effect of time can be modelled as a random process. Let $d(y, t|x)$ denote the probability of a state $\underline{y}$ being observed after elapsed time t from the initial state $\underline{x}$. Then the expected utility of choosing the alternative represented by a datum $\underline{x}^\delta$ with face value $\underline{x}$ and timestamp showing elapsed time t is

$$U_A(i, \underline{x}^\delta) = \Sigma_a \Sigma u(a(y))d(y, t|x) = \Sigma_y U_A(i, \underline{y})\, d(y, t|x). \qquad (6)$$

In many cases a datum remains valid to a cut-off point, after which it conveys no information at all. In other cases, there may be a steady decay in the probability of data being valid, so that $d(y, t|x)$ is a normal distribution about the mean given by the initial datum, with standard deviation proportional to the elapsed time.

7. The value of information

The previous sections considered the effect of content quality of information on decision making based on utility gained from satisfaction of a given set of requirements *A*. The next sections look more directly at the effect of content quality on the value to a recipient of the information provided by a source.

In this section we define the value of information in the special case that it correctly represents the state of the world.

Koamea developed a data valuation model which captures context through decision trees and quality through the effect of accuracy and availability on probabilites at branch points in the decision trees (Koamea 1994). His notion of information value was therefore tied to a certain decision making task structure. Our approach is more general.

Defining the value of information has always been controversial. The statistical definitions (Shannon and Weaver 1948) gives the information $I(\underline{x})$ in a datum $\underline{x}$ as

$$I(\underline{x}) = -\log(p(\underline{x})) \qquad (7)$$

where $p(\underline{x})$ is the probability of the datum. This definition does not allow for semantics. No-one has developed a robust or widely accepted way to model the role of the recipient and/or the task to which the information is applied (eg. Mackay 1970; Devlin 1992). However, researchers agree that information is associated with reduction in uncertainty about the world, and this is usually presented as the elimination of previously plausible alternative states of the world (eg. Hintakka 1969; Lozinskii 1994). The effect of uncertainty about what $\underline{x}$ conveys is to lower the value of information, since $\underline{x}$ eliminates fewer alternatives.

So along with most workers, including the above, we assume that there are a finite number of possible states of the world. The probability of a statement is the fraction of possible states of the world in which the statement holds (Lozinskii 1994). The usual information value is then obtained by applying *(7)*. However, even when the notion of "state of the world" is severely constrained, most of the things that determine a state of the world are not of any intererest to a particular person engaged in a task, and so this definition of information is not useful. How should the subjective assessments of the person and the task to which information is to be applied be modelled?

Consider a (price insensitive) diner selecting a restaurant on the basis of what it offers on the menu. She is interested in a set of possible meals, such as "Thai chicken soups with lots of chilli", "French fish main meals", "Italian desserts with marsala". Or consider the chef selecting meat for the day's catering. He is interested in a set of possible ingredients, such as "sides of beef of any size if they are less than $3 per kilo", "good quality fillet at less than $13 per kilo", "skinned chicken breasts at less than $5 per kilo".

The potential diner or the chef wants to know the relevance of an information site to their interests. The things of interest to them can be presented as a set of statements $D = \{f\}$. Not all interests are equally important. So the ranking of interests can be reflected in a weighting function u defined on the set of statements D. Now, data offered by a site are of value as information to the diner or the chef only if they *say something about* the statements of interest. For example, a diner interested only in vegetarian main courses is not interested in a menu consisting of meat dishes and desserts. Thus, the weighted sum of the statements which are implied by the information at a site indicates the relevance of that site to the person with these interests.

What a datum $\underline{x}$ says about a statement f is captured in the change in the information gained from knowing f is true with or without knowledge of $\underline{x}$. For example, if a diner learns that a restaurant is Thai, then there will be less information gained when she eventually learns that the restaurant offers Thai chicken soups with lots of chilli than there would have been had she not known whether the restaurant was French, Indian, or Thai. So the fact $\underline{x}$ that the restaurant is Thai increases the chance that the interest statement f ="Thai chicken soups with lots of chilli" will be satisfied by the restaurant. This change in information value is essentially the quantity $log(p(\phi)) - log\ (p(\phi|\underline{x})) = - log\ (p(\phi\,|\,\underline{x})/p(\phi))$ on which the next definition is based.

The information value of the datum $\underline{x}$ obtained from site i to a recipient interested in ϕ is defined as

$$I_\phi(\underline{x}) = - log\ (p(\phi\,|\,\underline{x})/p(\phi)), \qquad \text{if } p(\phi\,|\,\underline{x}) \geq p(\phi) \neq 0;$$

$$= - log\ (p(\phi) + (1 - p(\phi))\, p(\phi\,|\,\underline{x})\,|\,p(\phi)) \qquad \text{if } 0 \neq p(\phi\,|\,\underline{x}) < p(\phi);$$

$$else\ I_\phi(\underline{x}) = 0. \qquad (8)$$

See (Aisbett and Gibbon 1996) for details. The rationale for not simply taking the first of these expressions is as follows. If the number of states of the world compatible with ϕ *given* $\underline{x}$ is the same as the number compatible with ϕ before $\underline{x}$, then clearly no information about ϕ is transmitted by the datum $\underline{x}$. In any other circumstance, $I_\phi(\underline{x})$ should be positive because, whether or not the number of states compatible with ϕ increases or decreases as a result of receiving $\underline{x}$, information about ϕ has been conveyed. Clearly, $I_\phi(\underline{x})$ should not exceed $I(\phi)$, the value of information in ϕ itself. This maximum should however be attained either when $\underline{x}$ implies ϕ or when $\underline{x}$ implies not ϕ. In the latter case $I_\phi(\underline{x})$ is the value of the knowledge that ϕ is certainly false.

If the recipient has a set of interests D, then the information in $\underline{x}$ is the weighted sum of the information relative to an interest, over the statements in the set, viz,

$$I_D(\underline{x}) = \Sigma\{u(\phi)\, I_\phi(\underline{x}) : \phi \in D\}. \qquad (9)$$

8. The effect of content quality on the relative value of information

The last section defined a plausible value of information which took into account the interests of the recipient of a datum. This section considers the effect of content quality on this value. This

corresponds to the effect on the ability of a person to learn about certain subjects from the information in that datum.

Let $p_i(\underline{y}|\underline{x})$ be the conditional probability that a real world situation described by $\underline{y}$ is being presented by source i as $\underline{x}$ and again, let $\underline{x}^\delta$ denote the uncertain value presented as $\underline{x}$. The relative error in the value of information when a real world state $\underline{y}$ is presented as $\underline{x}$ is

$$\partial I_\phi(\underline{x}^\delta) = |1 - I_\phi(\underline{y})/I_\phi(x)|. \tag{10}$$

The computations to account for the various data quality parameters are then analogous to those in section 6, using equation *(10)* instead of equation *(1)*.

If a source has reliability level $r(s)$ and presents the information in $\underline{x}^\delta$ to the recipient interested in ϕ then the expected relative error in the information is

$$\partial I_\phi(\underline{x}^\delta) = (1-r(s)) \sum_y |1 - I_\phi(\underline{y})/I_\phi(x)| \, p'(\underline{y}) \tag{11}$$

where p' is the *a priori* distribution of occurrences of data values in the domain.

If $c(\underline{x}^\delta)$ is the credibility of an uncertain datum $\underline{x}^\delta$ for which the source reliability is less than 1, then the expected error is

$$\partial I_\phi(\underline{x}^\delta) = (1-r(s)\, c(\underline{x})) \sum_y |1 - I_\phi(\underline{y})/I_\phi(x)| \, p'(\underline{y}). \tag{12}$$

When accuracy is reported through the explicit or implicit probability distribution p'', the expected relative error in the value of information is

$$\partial I_\phi(\underline{x}^\delta) = \sum^{x+e}_{x-e} |1 - I_\phi(\underline{y})/I_\phi(x)| \, p''(\underline{y}). \tag{13}$$

Finally, if $d(y, t|x)$ denotes the probability of a state $\underline{y}$ being observed after elapsed time t from the initial state $\underline{x}$, then the expected relative error in the information after elapsed time t is

$$\partial I_\phi(\underline{x}^\delta) = \sum |1 - I_\phi(\underline{y})/I_\phi(x)| d(y, t|x). \tag{14}$$

9. Conclusion

This paper extended the standard decision making technique of maximising expected utility to quantitatively take into account labelling of data for traditional quality parameters, as used in military intelligence. We presented a new logical definition of the value of information without uncertainty, which took into account the task to which the information is to be applied. We then presented a quantitative measure of the relative error in the value of information conveyed when data quality is taken into consideration.

While quantitative estimates such as we have presented cannot hope to map accurately onto real world decision making, they can provide indications of the scale of effects, and so serve as a guide, or as a caution. In some organisational situations, policies may need to be formulated to specify acceptable compromises between the mandated use of in-house verified databases and allowing decision makers to base opinions on timely information from possibly unreliable sources. The actual policy in such situations may not be as important as the fact that the organisation has drawn attention to the data quality issue by formulating a policy in the first place.

As indicated earlier, neither the prescriptive view of decision theory nor any theory of information based on reduction of uncertainty has proved to be an adequate model of actual human processes. The effect of quality on decision making and perceived value of information is thus far more complex than suggested by our results. More descriptive research needs to be carried out in this important area, and there is an urgent need for agreed standards on data quality labelling. This is becoming critical as business decision making becomes increasingly decentralised and based on larger volumes of information obtained by individual decision makers from almost any available source.

10. References

Aisbett, J. and Gibbon, G. "A practical definition of the information in a logical theory". Submitted to Journal of Experimental and Theoretical Artificial Intelligence, 1996.

Bell, D., Raiffa, H. and Tversky, A. Decision Making: Cambridge University Press, 1988.

de Kleer, J. "An assumption based TMS", Artificial Intelligence, 28, 1986, pp 127-162.

Devlin, K. Information and Logic: Cambridge University Press, 1992.

Dvir, V and Evans, S. "A TQM Approach to the Improvement of Information Quality", 1996 Conference on Information Quality, Sloan School of Management, MIT, 1996.

Flickner, M. et al, "Query by image and video content: the QBIC system", IEEE Computer 28, 9, 1995, pp 23-32.

Hintakka, J. "On Semantic Information". In Hintikka, J. and Suppes, P. (Ed) Information and Inference: Riedel, Dordrecht, 1970.

ISO TC 211 International Standards Organisation Technical Committee 211, 15046-4 CEN/TC 287 N 466, Geographic information - Data description - Quality; WG 2. 1996.

Koamea, P "Valuation of data: a decision analysis approach", Total Data Quality Management Working Series, TDQM-94-09. Sloan School of Management, MIT, 1994.

Lozinskii, E. " Information and evidence in logic systems". Journal of Experimental and Theoretical Artificial Intelligence 6, 1994, pp 163-193.

Mackay, D. Information, mechanism and meaning: MIT Press, 1969.

Manual of Land Warfare. Combat Intelligence Pamphlet . Department of Defense, 1979.

March, J. and Shapira, Z. " Behavioural decision theory and organizational decision theory". In Ungson G. and Braunstein D. (eds) Decision making: an interdisciplinary approach: Wadsworth Inc, Belmont, 1982, pp 92-116.

Redman, T. "Improve data quality for competitive advantage", Sloan Management Review , 36, 2, 1995, pp 99-106.

Segev, A. "On Information Quality and the WWW Impact", 1996 Conference on Information Quality, Sloan School of Management, MIT, 1996..

Shannon, C.and Weaver, W. The mathematical theory of communication: University of Illinois Press, 1949.

Wang, R., Strong, D and Guarascio, L. "Data Consumers Perspectives Of Data Quality", Total Data Quality Management Working Series, TDQM-94-01. Sloan School of Management, 1994.

Wand, Y., and Wang, R. "Anchoring data quality decisions in ontological foundations". Total Data Quality Management Working Series, TDQM-94-03. Sloan School of Management, 1994.